A Statistical Model to Investigate the Reproducibility Rate Based on Replication Experiments

Francesco Pauli

DEAMS, University of Trieste, P.le Europa 1, 34127 Trieste, Italy
E-mail: francesco.pauli@deams.units.it

Summary

The reproducibility crisis, that is, the fact that many scientific results are difficult to replicate, pointing to their unreliability or falsehood, is a hot topic in the recent scientific literature, and statistical methodologies, testing procedures and *p*-values, in particular, are at the centre of the debate. Assessment of the extent of the problem–the reproducibility rate or the false discovery rate–and the role of contributing factors are still an open problem. Replication experiments, that is, systematic replications of existing results, may offer relevant information on these issues. We propose a statistical model to deal with such information, in particular to estimate the reproducibility rate and the effect of some study characteristics on its reliability. We analyse data from a recent replication experiment in psychology finding a reproducibility rate broadly coherent with other assessments from the same experiment. Our results also confirm the expected role of some contributing factor (unexpectedness of the result and room for bias) while they suggest that the similarity between original study and the replica is not so relevant, thus mitigating some criticism directed to replication experiments.

Key words: *p*-value; false discovery rate; reproducibility crisis; mixture model.

1 Introduction

A number of authors have recently cast doubt on the reliability of published scientific results, arguing that a high number of scientific findings are false or hardly reproducible (Leek & Peng, 2015b; 2015a; Nuzzo, 2014), the fact that a relevant 'crisis of reproducibility' is experienced nowadays is an opinion shared by the large majority of researcher interviewed by Nature in a recent survey (Baker, 2016).

These criticisms are directed at results based on a commonplace paradigm–or 'ritual' as Gigerenzer (2004) provocatively called it–which roughly goes as follows: a confirmation of a theory, meaning here the existence of a relationship between an outcome and a factor, is generally claimed whenever a null hypothesis of absence of the relationship is statistically rejected, the observed significance level (*p*-value) being a measure of the support given by the data to the theory. The central role of statistical hypotheses testing, and *p*-values in particular, has lead to pinpointing it as one of the main culprits (McCloskey, 1995; Goodman, 1999a; Cohen, 1994), reviving an old controversy (Krantz, 1999) and prompting the recent ASA statement on the correct use of *p*-values (Wasserstein & Lazar, 2016). Some scholars (Leek & Peng, 2015a; Gigerenzer, 2004) have argued that the problem lies with the paradigm itself rather than the

p-value specifically, although it is possible that the use of statistical testing as a way to draw conclusions may worsen things by masking phenomena due to researcher bias (loosely speaking the more or less conscious manipulation of data to obtain a given result, see Section 2). The statistical methodology is often seen both as part of the problem and as a mean to improve the situation by respondents to the Nature survey (Baker, 2016). In fact, a number of proposals have been put forward to substitute hypotheses testing with alternative methods, for example, Bayes factors (Goodman, 1999b) or measures of prediction error within a model selection framework (Reiss, 2015; Wagenmakers, 2007). At present, however, there is no clear consensus on whether these alternatives are preferable to hypotheses testing.

We will not discuss further the epistemological aspects of the issue or whether the p-value is to be blamed; rather, we focus on the empirical evidence concerning the reliability of results obtained by means of the paradigm, where by reliability, we refer to the false discovery rate (FDR) or the rate of reproducibility.

The idea of a FDR across science (or a discipline) is intuitively clear, and it is undoubtedly relevant to assessing reliability of 'science'. [It also dates back a long time, it is reported that the so-called 5σ threshold, which is used in physics as a critical level of significance, stems from a reasoning based on the number of hypotheses tested in the scientific literature at that time (1968) and the expected number of discoveries with a lower threshold (Demortier, 2010).] On the contrary, it is rather difficult to define such a FDR on formal grounds due to the difficulties in defining a population of scientific findings on which it should be measured and also to the difficulties in identifying the false ones. (It is also an ill-posed problem in statistical terms as hypotheses testing is not designed to control it; the discussion, however, goes beyond the statistical methodology, and so this consideration is insufficient to dismiss the issue.)

When experimental results are considered, the idea of a deceitful result is made more concrete by looking at lack of reproducibility, that is, whether, upon replication of the experiment, it leads to a different conclusion. The two concepts of falsity and lack of reproducibility of a result do not equate, as a different conclusion from a replica experiment may well be compatible with the conclusion of the first when it comes to statistical conclusions. However, it is reasonable to see lack of reproducibility as an indicator of reliability [if not a proxy for falsity (Prinz *et al.*, 2011; Begley & Ellis, 2012)] as is discussed more thoroughly in Section 3.

Despite the difficulties in defining the concepts, there have been various attempts at assessing the reliability of scientific literature either in terms of FDR (Section 2) or of rate of reproducibility (Section 3).

We will not discuss here the assessments not directly based on empirical evidence, that is, those where theoretical calculations are performed obtaining hypothetical FDRs based on some assumption on the proportion of false hypotheses tested and the procedure employed to confirm a discovery (Ioannidis, 2005b; Berger, 2003).

Empirical assessments stem from examining a set of results (usually from published papers belonging to some more or less specific field). Because the set of studies is typically (forcibly) chosen according to criteria that do not allow for generalisation beyond the set itself, finding a high FDR or low reproducibility does not necessarily indicate that the results in the field are unreliable. Still, it is a relevant information: at the very least, the conclusions apply to those results that have been examined and those may be relevant per se. Furthermore, a lack of reliability within the sample of results is suggestive of the fact that the paradigm used to provide those result may be problematic. In other words, we can not generalise the supposedly high FDR or low reproducibility rate to the field, but it may be evidence that the paradigm that has been used is capable of leading to a high FDR or low reproducibility rate.

Besides the generalisation problem, even drawing conclusions on the sample itself is not straightforward due to the aforementioned difficulties in identifying false results and the fact

that reproducibility is not clearly defined when statistical conclusions are involved. We review some proposals, and the issues involved, in Sections 2 and 3, the former being more focused on FDR assessments, the latter on rate of reproducibility. It is worth to note that none of these attempts at quantifying the extent of the issue is concerned with those disciplines where the number of variables dominates the sample size and one would expect the problem to be even worse. In fact, this is a difficult context in which the statistical techniques involved are slightly less conventional and, to some extent, do allow for multiple testing. However, many authors argue, qualitatively and anecdotally, that this is a relevant issue, for example, in genetics (Ioannidis *et al.*, 2001; Hunter & Kraft, 2007; Callaway, 2017; Eaves, 2006) and neuroimaging (Vul & Pashler, 2012; Vul *et al.*, 2009).

We then focus on analysis of reproducibility, in particular, we consider data from a study in which replication was attempted for a hundred published experiments in psychology (Open Science Collaboration, 2015), and we propose (in Section 4) a model for the experimental results, summarised by the respective p-values, to estimate the reproducibility rate and to assess the relevance of some characteristics of the original results and the replica that may affect reproducibility. Because a major difficulty in analysing this kind of data is specifying a probability distribution for p-values under the alternative hypotheses (and even under the null in certain circumstances as discussed in Section 4), we perform a simulation study (Section 5) to assess robustness of our model specification, also with respect to alternatives.

2 Empirical Evidence on the Reliability of Scientific Results

One of the purported reasons for the lack of reproducibility of published results is related to the fact that the original result (p-value) does not come from a single pre-specified test but is one (the lowest) among the p-values obtained by the experimenter who performs a number of tests exploring different theories/hypotheses. This phenomenon, called p-hacking (selective reporting), may easily lead to low p-values in the absence of a signal. Moreover, a similar phenomenon may occur even when a single, maybe pre-specified, hypothesis is tested, but the researcher has a number of degrees of freedom in defining the details of the analysis (e.g. inclusion/exclusion of variables or observations), and the choices are driven by the data. This was named the Garden of Forking Paths (GOFP) by Gelman and Loken (2014) (an example is given by the recoding of variables we did for the present analysis, see Table 2). Simonsohn *et al.* (2014) noted that the shape of the frequency distribution of the p-values from a sample of studies may allow to gain insights on the extent of phenomena such as p-hacking and GOFP and hence on the reliability. For example, a prevalence of p-values just below 0.05, which is coherent with the p-hacking and GOFP phenomena, has been observed within biology (Head *et al.*, 2015), economics (Brodeur *et al.*, 2016) and psychology (Masicampo & Lalande, 2012).

The approach of Simonsohn *et al.* (2014) is quite indirect, in terms of reliability assessment, in that they seek for patterns that reveal the existence of a problem. Jager and Leek (2014) take a more direct (and audacious) approach and employ methods from genomics to estimate the rate of false discoveries in medical journals. They collect a sample of 5 322 p-values less than 0.05 from five medical journals from 2010 to 2015 and model them according to a mixture of a uniform distribution in $[0, 0.05]$, which is the distribution of the p-value under the null hypotheses, and a beta distribution, truncated at 0.05, which is a plausible model, according to the literature, for the behaviour of a set of p-value under the alternative hypotheses. The estimate of the weight of the first component of the mixture ($14\% \pm 1\%$) is then taken as an estimate of the FDR. The task that Jager and Leek (try to) accomplish is rather difficult and ambitious, and a number of criticisms have been raised concerning their methods and assumptions (see the discussion of the cited paper). One relevant issue is that, in order to get a relatively big sample,

automatic text mining methods have been used to collect the p-values; thus, there is no distinction between primary and secondary findings, and when multiple p-values are scraped from the same paper, there is no control over correlation between them: moreover, and more generally, the criteria for inclusion in the study include typographical characteristic, which may bias the results in unpredictable ways (Goodman, 2014). Also, Gelman and O'Rourke (2014) point out that p-values have a uniform distribution under the null only in ideal settings and likely not when the p-values may suffer from phenomena of p-hacking or GOFP. It is also dubious whether a common distribution is adequate to describe the behaviour of the p-values when the null is false, because the sample includes p-values from an (uncontrolled) variety of situations with different signal strengths and different types of tests.

3 Assessing the Reliability through Replication

The idea of assessing the reliability of a finding by trying to reproduce it, that is, by replicating the experiment that lead to that finding, is straightforward and central to the scientific method, although there are very limited incentives for researchers to do so. On the other hand, a perfect replication of an experiment is seldom possible, even when enough information on materials and methods of the original experiment are given and both sufficient expertise and suitable resources are available (Schwalbe, 2016), so any conclusion should allow for the non-perfect replica (Kaiser, 2015; Yong, 2012). Furthermore, even if a perfect replica is performed, inasmuch as the results of the original studies are given in statistical terms, that is, the theory under scrutiny does not offer a precise prediction of the experimental results, it is not clear what a failure to reproduce the conclusion means, as the fact that a second study accepts the (same) null is not necessarily a contradiction (Simonsohn *et al.*, 2013) because 'The difference between *significant* and *not significant* is not itself statistically significant' (Gelman & Stern, 2006). Notwithstanding the inherent difficulties, replicating a sample of experiments in a field may shed light on the reliability of discoveries from that field or on the characteristics of experiments which may affect reliability.

In a paper that sparkled the recent debate, Ioannidis (2005a) considered 49 highly cited clinical research studies and sought for replications in the literature: of the 46 studies claiming an effect, 11 were unchallenged, 7 were contradicted and for other 7, the initial estimate of effect was reduced. Other authors have discussed the rate of reproducibility in medical science finding rate of confirmation as low as 11% (Begley & Ellis, 2012), although in pre-clinical research, which is a difficult area; 20 – 25% in the pharmaceutical context [where false discoveries may also have financial consequences when the objective is to evaluate a pharmaceutical company based (also) on the research behind its strategy (Osherovich, 2011)].

More pro-actively, initiatives aimed at trying to replicate existing studies are in place in medicine both by pharmaceutical companies and by academic researchers (Kaiser, 2015; Errington *et al.*, 2014; Mobley *et al.*, 2013), in economics (Camerer *et al.*, 2016) and in psychology (Klein *et al.*, 2014; Open Science Collaboration, 2015). The OSC (Open Science Collaboration, 2015) considered a (more or less) random sample of papers from three leading journals (one general purpose and two discipline-specific), within the cognitive and social-personality sub-disciplines. The selection of papers was driven by a compromise between generalisability of reproducibility estimate and the feasibility of the replica experiments. For each study, the OSC attempted reproduction of the (main) result. Each experiment was replicated by testing a single, predetermined, effect, thus leaving no (or very little) room to p-hacking or GOFP phenomena. Observed significance levels of the original studies and the corresponding replicas are reported in Figure 1. The relationship between the two p-values appears relatively weak unless the p-values are very low (well below the 0.05 threshold). As

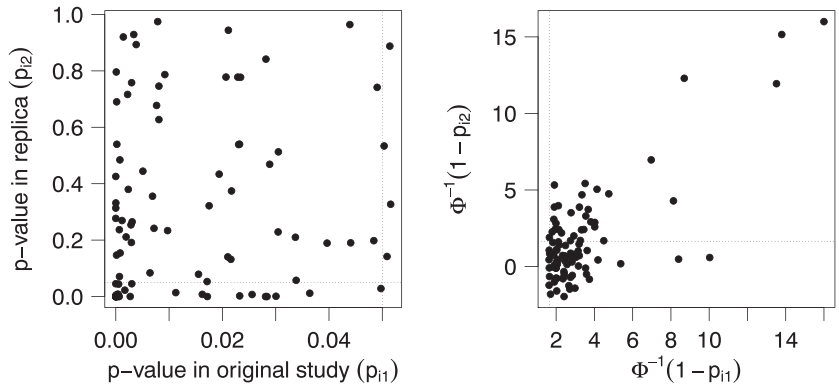

Figure 1. *Open Science Collaboration replication experiment: p-values from original studies and replicas compared (94 studies) on natural (left panel) and transformed scale (right panel).*

Table 1. *Assessment of reproducibility rate within the OSC study.*

Number of studies	100
Replications with significant results	36%
Original effect size within 95% CI of the replica	47%
Subjectively rated as successful replications	39%
Significant combining data from original study and replica	68%

CI, confidence interval; OSC, Open Science Collaboration.

Table 2. *Additional informations on the studies replicated within the OSC analysis* (Open Science Collaboration, 2015)*, original scale and modified scale for the present analysis.*

Description (name in model)	Values		
	Original	(Freq)	Modified
Opportunity for researcher expectations to influence results (b)	No	(45)	1
	Slight	(32)	2
	Moderate	(14)	3
	Strong	(3)	
Surprising result* (r)	1	(3)	1
	2	(24)	
	3	(29)	2
	4	(33)	3
	5	(5)	
Similarity between original and replica (s)	Some what similar	(4)	1
	Moderately similar	(8)	
	Very similar	(22)	2
	Extremely similar	(38)	3
	Virtually identical	(22)	4

* Originally on a continuous scale, rounded to the nearest integer as a first step.

outlined previously, it is not straightforward to determine if a replication is successful, so the authors report different plausible evaluations of the reproduction rate that range from 36% to 68% (Table 1).

The OSC inquiry has been criticised on the grounds that the replications may differ from the original studies because the sample in the replication study was drawn from a different population (e.g. a study on Americans was replicated drawing a sample from Italian population) or because the procedure was different (e.g. a different stimuli was considered), and these differences may decrease the rate of successful replication (Gilbert *et al.*, 2016). This is admittedly (Anderson *et al.*, 2016) a limitation of any analysis based on replication, and there is an obvious trade-off between the accuracy with which each experiment can be replicated and the number of experiments that can be replicated. An attempt at mitigating this issue was in fact made by considering an indicator of similarity between the original study and the replica (see Table 2).

Furthermore, Gilbert *et al.* (2016) argue that performing a single attempt at reproducing the results of a study amounts to a low power procedure. While it is intuitively true that a low power replica is less likely to reproduce an actual signal, it is unclear whether using a more powerful technique in the replication is susceptible to lead to a significantly higher reproducibility rate (Anderson *et al.*, 2016). Ongoing replication projects such as Klein *et al.* (2014, 2015) may help shedding more light on this issue.

Finally, a major limit of the assessments of reliability, fully acknowledged by the OSC, is that it is not clear to what extent the results can be generalised; the population is poorly defined, and the sampling strategy is far from being a random sample.

4 A Bayesian Mixture Model for the Open Science Collaboration Studies

Despite the limitations that an inquiry such as that of the OSC may have, we believe that it contains relevant information, and so we discuss possible modelling strategies to learn from the data that were produced. We propose a model for the p-values arising from the original results and the replicas with the aim of assessing the reproducibility rate and also of investigating whether some characteristics of the studies are associated with how likely they are to reproduce. In particular, we use a mixture model, similar to the one proposed by Jager and Leek (2014), in which each pair of p-values (original and replica) comes from a mixture distribution were one component describes the p-value behaviour under the null hypothesis while the second corresponds to it being false. Because the original studies all claimed a significant result, the weight given to the second component of the mixture can be seen as a reproducibility rate.

Let then p_{ij} represent the outcome (p-value) for experiment $i = 1, \ldots, n$ in the original study ($j = 1$) and in the replication ($j = 2$). We transform the p-values on a Gaussian scale by letting $y_{ij} = \Phi^{-1}(1 - p_{ij})$ and model y_{ij} as a mixture of a $N(0, 1)$ with probability $(1-\theta)$ and a $N(\mu_i, 1)$ with probability θ, both truncated for y_{i1} to allow for the fact that y_{i1} is censored because we consider only studies for which the p-value is below a certain threshold. The density for the pair (y_{i1}, y_{i2}) is then

$$f(y_{i1}, y_{i2}|\theta, \boldsymbol{\mu}) = (1-\theta)\frac{\phi(y_{i1})}{p^*}\phi(y_{i2}) + \theta\frac{\phi(y_{i1}-\mu_i)\phi(y_{i2}-\mu_i)}{1-\Phi(\Phi^{-1}(1-p^*)-\mu_i)}, \tag{1}$$

for $i = 1, \ldots, n$, where $p^* = \max\{p_{i1} : i = 1, \ldots, n\}$ is the maximum p-value among the studies of the sample. We also assume that $\mu_i \sim N(\lambda, \nu^2)$ and $\lambda, \nu \sim hN(0, 100)$ (half normal with parameters 0, 100), $\theta \sim \text{Unif}(0, 1)$. Estimation is performed using STAN (Carpenter *et al.*, 2017; Stan Development Team, 2016) within R (R Core Team, 2015), and results are based on four parallel chains of length 5 000.

We note in pass that the model assumes that the distribution of the p-value is the same in the original experiment and the replica (but for the truncation), which is reasonable because in the OSC study, the experiments were repeated in the same settings; however, it would be possible to generalise the model to allow for differences, for instance, if the original and replica had different power, this could be allowed for by specifying a different variance in the two distributions of the second part of the mixture.

The first component of the mixture, which assumes a uniformly distributed p-value, describes the behaviour of the test when the null hypothesis is true in the ideal situation where no p-hacking or GOFP are in place. In principle, these assumptions are reasonably tenable for the replication but are disputable for the test performed in the original study. The second component assumes some deviation (expressed by the parameter μ_i) from the null. This is a broad approximation, because the testing statistics involved in the studies are different and do not have the same behaviour under the alternative (see the discussion on this issue in Section 5). These considerations make the parameters μ_i, λ, ν difficult to interpret.

Keeping into account the aforementioned limits, one may interpret the parameter θ as the rate of successful replication. The posterior distribution for θ is reported in the first panel of Figure 2 and suggests a value between 50% and 60%. With respect to the conclusion of the OSC, this figure may be loosely compared with the rate of significant results obtained by combining the original and replicated studies (last line of Table 1).

The OSC data also include some characteristics of the studies that may be associated to their reliability, in particular, we consider three of them, all subjectively rated by the OSC researcher, reported in Table 2. The opportunity for researcher expectations to influence results (b_i) is related to the extent to which p-hacking or GOFP may affect the test result in the original study. A surprising result (r_i) may be seen as less likely to be true. Finally, a lack of similarity between the original study and the replication (s_i) may introduce random error and lead to a lower reproducibility rate (Gilbert *et al.*, 2016).

In order to explore to what extent these characteristics of the studies affect reproducibility, we specify a model in which the parameter θ_i is allowed to vary depending on them. Let then

$$f(y_{i1}, y_{i2}|\theta, \boldsymbol{\mu}) = (1-\theta_i)\frac{\phi(y_{i1})}{p^*}\phi(y_{i2}) + \theta_i \frac{\phi(y_{i1}-\mu_i)\phi(y_{i2}-\mu_i)}{1-\Phi(\Phi^{-1}(1-p^*)-\mu_i)}, \tag{2}$$

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \alpha + \beta_{b_i} + \gamma_{r_i} + \delta_{s_i}, \tag{3}$$

where the triplet $\boldsymbol{\beta}$ has a prior distribution such that $\boldsymbol{\beta}/\sum\beta_j$ is uniformly distributed on $\{\boldsymbol{\beta} : \sum\beta_j = 0\}$ and $\sum\beta_j \sim N(0, 10)$. In other words, we impose a sum-to-zero constraint on

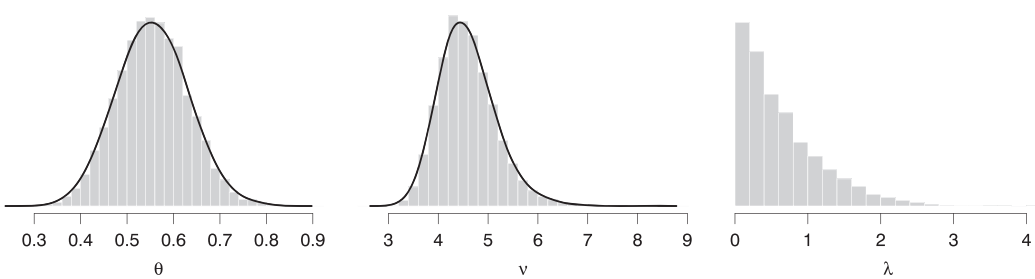

Figure 2. *Posterior distributions for θ, ν, λ.*

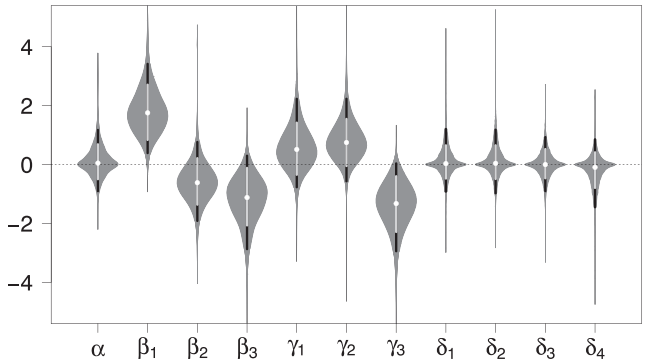

Figure 3. *Posterior distributions for the coefficients α, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$, segments represent 95% and 80% high posterior density credibility intervals, white dots represent the medians.*

the β_i. The same prior distribution is specified for the triplet $\boldsymbol{\gamma}$ and the quadruplet $\boldsymbol{\delta}$. Finally, $\alpha \sim N(0, 10)$.

The posterior distributions for the coefficients (Figure 3) show that a low opportunity for experimenter bias leads to a higher θ, while a more surprising result is associated to a lower θ. The degree of similarity between the original study and the replica seems not to affect the mixing parameter.

5 Model Robustness

We discuss three of our model assumptions that may be deemed dubious: the uniform distribution may not describe the behaviour of the p-value under the null if some form of p-hacking is present (as it is possible in the original experiment); p-values arising from different model structures may have different behaviours under the alternative hypotheses while we assume a common distribution (differing only because of the mean); and normality of $\Phi^{-1}(1 - p)$ under the alternative is not necessarily the optimal choice.

Different specifications could be used both under the null and under the alternative.
The assumption of a uniform distribution under the null does not allow for p-hacking/GOFP phenomena; however, the very nature of such phenomena, which are not precisely defined, makes it difficult to envisage an alternative distributional assumption to allow for them.

As far as the distribution of the statistics under the alternative hypotheses is concerned, we already mentioned in Section 4 that different test statistics may show different behaviours under the alternative. In general, a sensible option would be to use a more flexible specification for their distribution, for example, Jager and Leek (2014) mention the possibility of using a mix-ture of Beta distributions, but a non-parametric specification could also be used. An alternative strategy would be to model the differences, that is, diversify the second component depending on the test statistics. In the present analysis, the relative paucity of the sample and the fact that it involves a wide range of different test statistics (the studies in the OSC included analysis of variance with various designs, analysis of covariance, t-tests, χ^2 tests, correlation and regres-sion models with various specifications) deter us from employing the aforementioned strategies: the sample is not big enough to estimate a more flexible model, and differentiating with respect to the test statistic would require a non-trivial assessment of which groups may be described by a common distribution.

Our approach is then to assess robustness of our results to the violations of these assumptions by means of a small simulation study. Our aim is to simulate a sample of (pairs of)

p-values using a mechanism that resembles the way in which p-values are produced from actual experiments and that does not guarantee the validity of the aforementioned assumptions.

A pair of simulated p-values (original and replication experiments) is obtained by simulating two independent samples according to a given model and performing a test on both. The simulation, either under the null hypotheses or under a model within the alternative, is performed conditional on the first p-value being below 0.05. In order to mimic the p-hacking phenomenon, the simulations under the null were performed drawing the p-value from the original experiment as a minimum of 10 uniformly distributed p-values. (Note that the aforementioned procedure may be simplified by simulating directly the test statistics or even the p-value when its distribution is known under the null or alternative, the procedure is explained here in an easily generalisable form.)

We simulate samples of p-value pairs repeating the aforementioned procedure with varying proportions of true nulls. Within each sample, different data generating mechanisms are used in order to allow for the variability in the p-value distribution under the alternative hypothesis. In particular, in the simulations, we consider analysis of variance models with varying group numbers, sizes and signal strength.

On each sample, we estimate model (1) as well as two alternative specifications for the p-value when the null is false: a beta distribution with parameters a and b as in Jager and Leek (2014) and a Gaussian distribution for $-\log(p\text{-value})$ as suggested by Cox (2014) (see also Boos & Stefanski, 2012).

The posterior distributions of θ obtained using the three alternative specifications are compared with the true proportion of null hypotheses in Figure 4, all models perform reasonably well, with a slight advantage for the Gaussian specification.

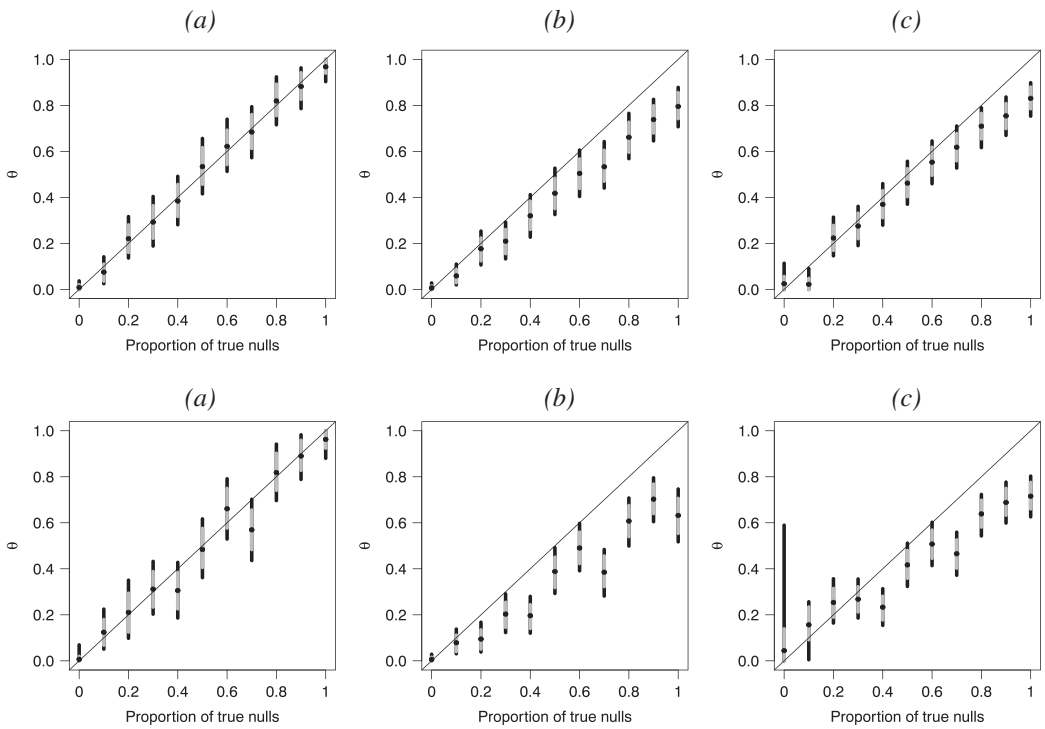

Figure 4. *Posterior distribution of the mixing parameter under different scenarios: (a) Gaussian distribution of* $\Phi^{-1}(p)$*; (b) Beta distribution for* p*; (c) Gaussian distribution for* $\log(p)$*; top row: average group size 100 observations, bottom row: average group size 10 observations.*

6 Discussion

Replication experiments such as the Open Science Collaboration (2015) one can be a valuable tool to investigate the reliability of scientific results and of the paradigm, which is used to reach them. However, it is not obvious how to interpret the data coming from such experiments. We proposed a model to estimate the reproducibility rate for the Open Science Collaboration (2015) sample of studies and to assess the effect of some characteristics of the studies on their reliability (as measured by reproducibility of results). Our analysis confirms that the more room there is for experimenter bias to influence outcomes, the less reproducible are the results, thus suggesting that p-hacking may in fact hinder the reliability of scientific conclusions. Moreover, the fact that the closeness of the replication to the original study have no association to the reproduction rate weakens some objections that have been moved against replication experiments (at least as far as the OSC study under scrutiny is concerned and keeping in mind that the closeness is subjectively rated).

Acknowledgements

I am grateful to the Associate Editor and an anonymous reviewer for their comments and suggestions. I would also like to thank Nicola Torelli for his comments on an earlier version of the manuscript.

References

Anderson, C. J., Bahník, Š., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R., Cheung, F., Christopherson, C. D., Cordes, A., Cremata, E. J., Della Penna, N., Estel, V., Fedor, A., Fitneva, S. A., Frank, M. C., Grange, J. A., Hartshorne, J. K., Hasselman, F., Henninger, F., van der Hulst, M., Jonas, K. J., Lai, C. K., Levitan, C. A., Miller, J. K., Moore, K. S., Meixner, J. M., Munafò, M. R., Neijenhuijs, K. I., Nilsonne, G., Nosek, B. A., Plessow, F., Prenoveau, J. M., Ricker, A. A., Schmidt, K., Spies, J. R., Stieger, S., Strohminger, N., Sullivan, G. B., van Aert, R. C. M., van Assen, M. A. L. M., Vanpaemel, W., Vianello, M., Voracek, M. & Zuni, K. (2016). Response to comment on "estimating the reproducibility of psychological science". *Science*, **351**(6277), 1037–1037.

Baker, M. (2016). Is there a reproducibility crisis. *Nature*, **533**, 452–454.

Begley, C. G. & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, **483**(7391), 531–533.

Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing. *Statist. Sci.*, **18**(1), 1–12.

Boos, D. D. & Stefanski, L. A. (2012). P-value precision and reproducibility. *Amer. Statist.*, **65**, 213–221.

Brodeur, A., Lé, M., Sangnier, M. & Zylberberg, Y. (2016). Star wars: The empirics strike back. *Amer. Econ. J.: Appl. Econ.*, **8**(1), 1–32.

Callaway, E. (2017). New concerns raised over value of genome-wide disease studies. *Nature News*, **546**(7659), 463.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M. & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, **351**(6277), 1433–1436.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P. & Riddell, A. (2017). Stan: A probabilistic programming language. *J. Stat. Softw.*, **76**(1), 1–32.

Cohen, J. (1994). The earth is round ($p < .05$). *Amer. Psychol.*, **49**, 997–1003.

Cox, D. R. (2014). Discussion: Comment on a paper by Jager and Leek. *Biostatistics*, **15**(1), 16–18.

Demortier, L. (2010). Dealing with data: Signals, backgrounds, and statistics. In *The Dawn of the LHC Era: TASI 2008: Proceedings of the 2008 Theoretical Advanced Study Institute in Elementary Particle Physics*, pp. 305. Boulder, Colorado, USA: World Scientific.

Eaves, L. J. (2006). Genotype× environment interaction in psychopathology: Fact or artifact. *Twin Res. Human Genet.*, **9**(1), 1–8.

Errington, T. M., Iorns, E., Gunn, W., Tan, F. E., Lomax, J. & Nosek, B. A. (2014). An open investigation of the reproducibility of cancer biology research. *Elife*, **3**, e04333.

Gelman, A. & Loken, E. (2014). The statistical crisis in science. *American Scientist*, **102**, 460–465.

Gelman, A. & O'Rourke, K. (2014). Discussion: Difficulties in making inferences about scientific truth from distributions of published p-values. *Biostatistics*, **15**(1), 18–23.

Gelman, A. & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *Amer. Statist.*, **60**(4), 328–331.

Gigerenzer, G. (2004). Mindless statistics. *J. Socio-Econ.*, **33**(5), 587–606.

Gilbert, D. T., King, G., Pettigrew, S. & Wilson, T. D. (2016). Comment on "estimating the reproducibility of psychological science". *Science*, **351**(6277), 1037–1037.

Goodman, S. N. (1999a). Toward evidence-based medical statistics. 1: The p value fallacy. *Ann. Internal Med.*, **130**(12), 995–1004.

Goodman, S. N. (1999b). Toward evidence-based medical statistics. 2: The Bayes factor. *Ann. Internal Med.*, **130**(12), 1005–1013.

Goodman, S. N. (2014). Discussion: An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*, **15**(1), 23–27.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T. & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biol.*, **13**(3), e1002106.

Hunter, D. J. & Kraft, P. (2007). Drinking from the fire hose–statistical issues in genomewide association studies. *N. Engl. J. Med.*, **357**(5), 436–439.

Ioannidis, J. P. (2005a). Contradicted and initially stronger effects in highly cited clinical research. *J. Amer. Med. Assoc.*, **294**(2), 218–228.

Ioannidis, J. P. (2005b). Why most published research findings are false. *PLoS Med.*, **2**(8), e124.

Ioannidis, J. P., Ntzani, E. E., Trikalinos, T. A. & Contopoulos-Ioannidis, D. G. (2001). Replication validity of genetic association studies. *Nature Genet.*, **29**(3), 306–309.

Jager, L. R. & Leek, J. T. (2014). An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*, **15**, 1–12.

Kaiser, J. (2015). The cancer test. *Science*, **348**(6242), 1411–1413.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B. Jr, Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W.E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., Hasselman, F. W., Hicks, J. A., Hovermale, J. F., Hunt, S. J., Huntsinger, J. R., IJzerman, H., John, M. S., Joy-Gaba, J. A., Kappes, H. B., Krueger, L. E., Kurtz, J., Levitan, C. A., Mallett, R. K., Morris, W. L., Nelson, A. J., Nier, J. A., Packard, G., Pilati, R., Rutchick, A. M., Schmidt, K., Skorinko, J. L. M., Smith, R., Steiner, T. G., Storbeck, J., Van Swol, L. M., Thompson, D., van't Veer, A. E., Vaughn, L. A., Vranka, M., Wichman, A. L., Woodzicka, J. A. & Nosek, B. A. (2014). Investigating variation in replicability. *Social Psychol.*, **45**(3), 142–152.

Klein, R. A., Vianello, M., Hasselman, F., Alper, S., Aveyard, M., Axt, J. & Nosek, B. (2015). Many labs 2: Investigating variation in replicability across sample and setting.

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *J. Amer. Statist. Assoc.*, **94**(448), 1372–1381.

Leek, J. T. & Peng, R. D. (2015a). Statistics: P-values are just the tip of the iceberg. *Nature*, **520**(7549), 612.

Leek, J. T. & Peng, R. D. (2015b). What is the question. *Science*, **347**(6228), 1314–1315.

Masicampo, E. J. & Lalande, D. R. (2012). A peculiar prevalence of p-values just below.05. *Q. J. Exp. Psychol.*, **65**(11), 2271–2279.

McCloskey, D. (1995). The insignificance of statistical significance. *Sci. Amer.*, **272**, 32–33.

Mobley, A., Linder, S. K., Braeuer, R., Ellis, L. M. & Zwelling, L. (2013). A survey on data reproducibility in cancer research provides insights into our limited ability to translate findings from the laboratory to the clinic. *PLoS ONE*, **8**(5), e63221+.

Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature*, **506**(7487), 150–152.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, **349**(6251), aac4716+.

Osherovich, L. (2011). Hedging against academic risk. *Sci.-Bus. eXchange*, **4**.

Prinz, F., Schlange, T. & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets. *Nature Rev. Drug Discov.*, **10**(9), 712.

R Core Team. (2015). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Reiss, P. T. (2015). Cross-validation and hypothesis testing in neuroimaging: An irenic comment on the exchange between friston and lindquist et al. *NeuroImage*, **116**, 248–254.

Schwalbe, M. (2016). *Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results: Summary of a Workshop.* Washington, DC: National Academies Press.

Simonsohn, U., Nelson, L. D. & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *J. Exp. Psychol. Gen.*, **143**(2), 534–547.

Simonsohn, U., Simmons, J. & Nelson, L. (2013). Anchoring is not a false-positive: Maniadis, Tufano, and List's (2014) 'failure-to-replicate' is actually entirely consistent with the original.

Stan Development Team. (2016). RStan: The R interface to Stan. R package version 2.14.1.

Vul, E., Harris, C., Winkielman, P. & Pashler, H. (2009). Puzzlingly high correlations in fmri studies of emotion, personality, and social cognition. *Perspect. Psychol. Sci.*, **4**(3), 274–290. PMID: 26158964.

Vul, E. & Pashler, H. (2012). Voodoo and circularity errors. *Neuroimage*, **62**(2), 945–948.

Wagenmakers, E.-J. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bull. Rev.*, **14**(5), 779–804.

Wasserstein, R. L. & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *Amer. Statist.*, **70**(2), 129–133.

Yong, E. (2012). Replication studies: Bad copy. *Nature*, **485**(7398), 298–300.